Lab 10: Linear Regression

Yutong Shao

Linear Regression Applied to Swimming

- We will consider Exercise 9.1 in Hoff very closely to illustrate linear regression.
- The data set we consider contains times (in seconds) of four high school swimmers swimming 50 yards.
- There are 6 times for each student (four students in total), taken **every two weeks.**
- Each row corresponds to a swimmer and a higher column index indicates a later date.

```
library(ggplot2)
library(latex2exp)
library(MASS)
library(patchwork)
```

```
##
## Attaching package: ’patchwork’

## The following object is masked from ’package:MASS’:
##
##     area
```

Full conditionals (Task 1)

We will fit a separate linear regression model for each swimmer, with swimming time as the response and week as the explanatory variable. Let $Y_i \in \mathbb{R}^6$ be the 6 recorded times for swimmer $i = 1, 2, 3, 4$. Let

$$X_i = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ ... & \\ 1 & 9 \\ 1 & 11 \end{bmatrix}$$

be the design matrix for swimmer $i = 1, 2, 3, 4$. Then we use the following linear regression model:

$$\begin{aligned} Y_i \mid \beta_i, \tau_i &\sim \mathcal{N}_6\left(X\beta_i, \tau_i^{-1}\mathcal{I}_6\right) \\ \beta_i &\sim \mathcal{N}_2\left(\beta_0, \Sigma_0\right) \\ \tau_i &\sim \text{Gamma}(a, b). \end{aligned}$$

Derive full conditionals for β_i and τ_i. Assume that β_0, Σ_0, a, b are known.

Solution (Task 1)

The conditional posterior for β_i is multivariate normal with

$$\mathbb{V}[\beta_i \,|\, Y_i, X_i, \tau_i] = (\Sigma_0^{-1} + \tau X_i^T X_i)^{-1}$$
$$\mathbb{E}[\beta_i \,|\, Y_i, X_i, \tau_i] = (\Sigma_0^{-1} + \tau_i X_i^T X_i)^{-1}(\Sigma_0^{-1}\beta_0 + \tau_i X_i^T Y_i).$$

while

$$\tau_i \,|\, Y_i, X_i, \beta \sim \text{Gamma}\left(a + 3\,,\, b + \frac{(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)}{2}\right).$$

These can be found/verified in in Hoff in section 9.2.1.

Task 2

Complete the prior specification by choosing a, b, β_0, and Σ_0. Let your choices be informed by the fact that times for this age group tend to be between 22 and 24 seconds.

Solution (Task 2)

Choose $a = b = 0.1$ so as to be somewhat uninformative.

Choose $\beta_0 = [23 \;\; 0]^T$ with

$$\Sigma_0 = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}.$$

This centers the intercept at 23 (the middle of the given range) and the slope at 0 (so we are assuming no increase) but we choose the variance to be a bit large to err on the side of being less informative.

Gibbs sampler (Task 3)

Code a Gibbs sampler to fit each of the models. For each swimmer i, obtain draws from the posterior predictive distribution for the time of swimmer i if they were to swim two weeks from the last recorded time.

```r
data <- read.table("swim.dat", header=FALSE)
```

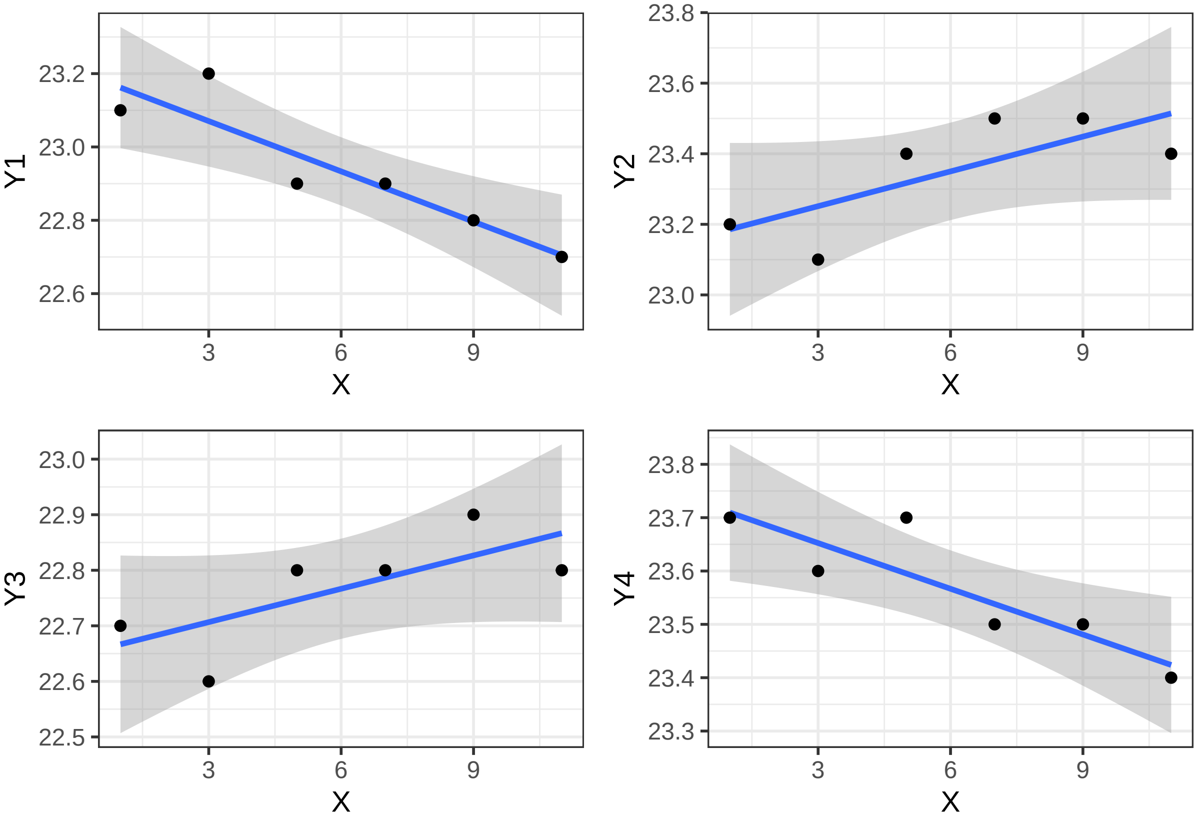
```
## Warning in read.table("swim.dat", header = FALSE): incomplete final line found
## by readTableHeader on 'swim.dat'
```

```r
df <- data.frame(t(data), X=c(1, 3, 5, 7, 9, 11))
colnames(df) <- c("Y1", "Y2", "Y3", "Y4", "X")



plot1 <- ggplot(df, aes(x=X,y=Y1))+
    stat_smooth(method=lm,formula="y~x")+
    geom_point()+
    theme_bw()
plot2 <- ggplot(df, aes(x=X,y=Y2))+
    stat_smooth(method=lm,formula="y~x")+
    geom_point()+
    theme_bw()
plot3 <- ggplot(df, aes(x=X,y=Y3))+
    stat_smooth(method=lm,formula="y~x")+
    geom_point()+
    theme_bw()
plot4 <- ggplot(df, aes(x=X,y=Y4))+
    stat_smooth(method=lm,formula="y~x")+
    geom_point()+
    theme_bw()


(plot1 | plot2) / (plot3 | plot4)
```



Gibbs Sampler

```
# frequentist approach
Y1 <- df$Y1
Y2 <- df$Y2
Y3 <- df$Y3
Y4 <- df$Y4
X <- df$X
n <- 6
```

```r
# Xi is a 6*2 matrix with the first column being ones
Xi <- cbind(rep(1,n), X)

# priors
beta0 <- c(23, 0)
sigma0 <- matrix(c(5, 0, 0, 2), nrow = 2, ncol = 2)
a <- b <- 0.1

BETA <- TAU <- NULL
tau <- rgamma(1, shape=a, rate=b)
bet <- beta0


Gibbs.sampler <- function(Y){

  for(s in 1:10000){

  # update beta
  inv.sig <- solve(sigma0)
  var.beta <- solve(inv.sig + tau * t(Xi) %*% Xi)
  mean.beta <- var.beta %*% (inv.sig %*% beta0 + tau * t(Xi) %*% Y)
  bet <- mvrnorm(1, mean.beta, var.beta)

  # update sigma
  an <- a + 3
  bn <- b + 0.5 * t(Y - Xi %*% bet) %*% (Y - Xi %*% bet)
  tau <- rgamma(1, shape=an, rate=bn)

  # store values
  BETA <- rbind(BETA, bet)
  TAU <- c(TAU, tau)
  }
  return(list(BETA = BETA, TAU = TAU))
}
```

```r
# posterior predictive distribution
PPD <- function(BETA, TAU){
  n <- dim(BETA)[1]
  Y_star <- NULL
  for (t in 1:n) {
    mu <- c(1,13) %*% BETA[n,]
    y <- rnorm(1, mean = mu, sd = 1/sqrt(TAU[n]))
    Y_star <- rbind(Y_star, y)
  }
  return(Y_star)
}
```

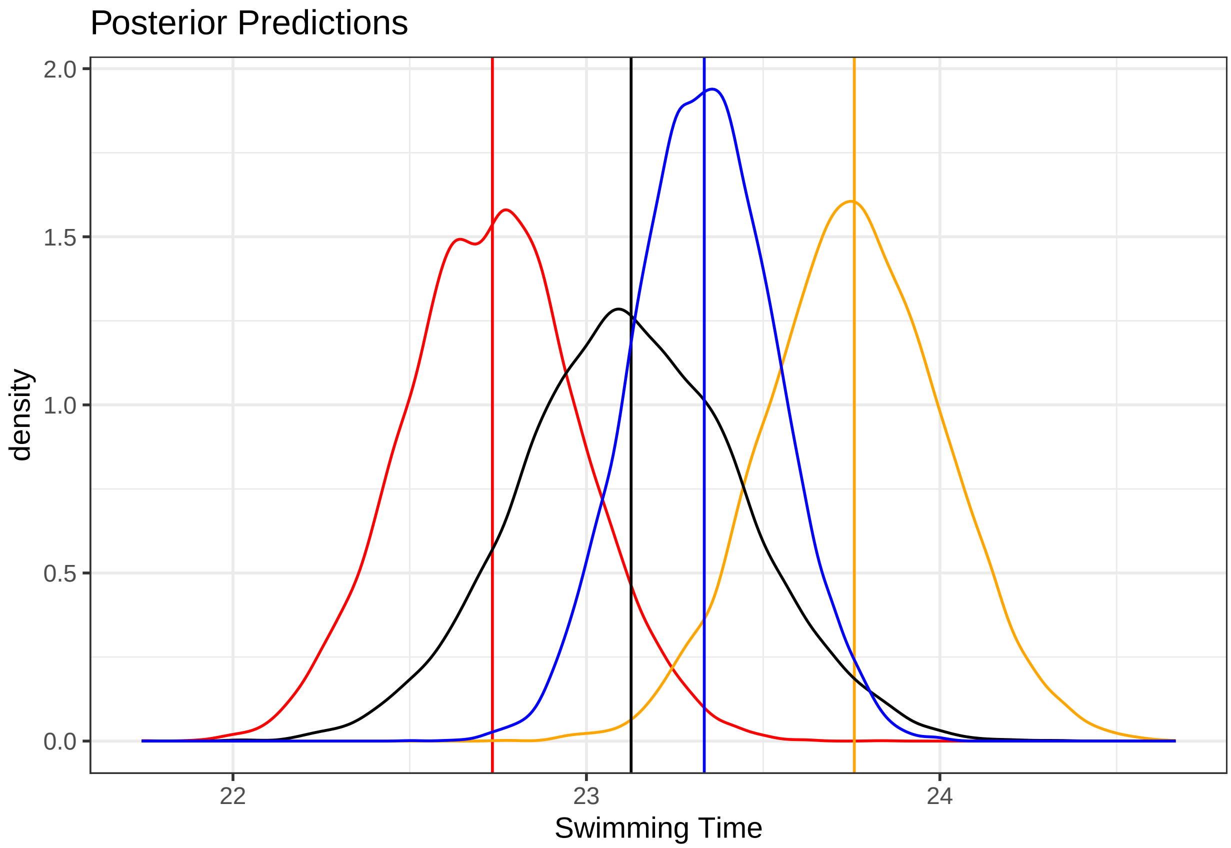
```r
set.seed(2023)
RES1 <- Gibbs.sampler(Y1)
pred1 <- PPD(RES1$BETA, RES1$TAU)

RES2 <- Gibbs.sampler(Y2)
pred2 <- PPD(RES2$BETA, RES2$TAU)
```

```
RES3 <- Gibbs.sampler(Y3)
pred3 <- PPD(RES3$BETA, RES3$TAU)

RES4 <- Gibbs.sampler(Y4)
pred4 <- PPD(RES4$BETA, RES4$TAU)
```

```
# plot posterior densities
ggplot()+
  geom_density(aes(x=pred1),col="red")+
  geom_vline(xintercept=mean(pred1),col="red")+
  geom_density(aes(x=pred2),col="orange")+
  geom_vline(xintercept=mean(pred2),col="orange")+
  geom_density(aes(x=pred3),col="black")+
  geom_vline(xintercept=mean(pred3),col="black")+
  geom_density(aes(x=pred4),col="blue")+
  geom_vline(xintercept=mean(pred4),col="blue")+
  ggtitle("Posterior Predictions")+
  xlab("Swimming Time")+
  theme_bw()
```



Posterior Prediction (Task 4)

The coach has to decide which swimmer should compete in a meet two weeks from the last recorded time. Using the posterior predictive distributions, compute $\Pr\left\{y_i^* = \max\left(y_1^*, y_2^*, y_3^*, y_4^*\right)\right\}$ for each swimmer i and use these probabilities to make a recommendation to the coach.

```r
N <- 10000
pred_mat <- data.frame(pred1, pred2, pred3, pred4)
freq <- apply(pred_mat, 1, which.max)

table(freq) / N
```

```
## freq
##      1      2      3      4
## 0.0009 0.8672 0.0502 0.0817
```

The first swimmer has the lowest probability of being the slowest, while the second swimmer is most likely to be the slowest. So, the coach should avoid sending the second swimmer to the competition and choose the first one.